IT550: Information Retrieval

Assignment 7: Text Classification

Instructor: Prasenjit Majumder

Learning Outcome: At the end of this assignment you will learn text classification using SVM. And also using micro and macro precision for evaluating your classifier.

1 Problem description

Text classification is the task of assigning a text to its proper category. The two main techniques that are used for text classification are supervised and unsupervised. In supervised learning we use labelled data to train our classifier, and then testing is performed. In unsupervised learning text clustering is performed. Text clustering is the task of grouping a set of unlabeled texts in such a way that texts in the same group (called a cluster) are more similar to each other than to those in other clusters.

2 Implementation

2.1 Dataset

- For this assignment we will use Jigsaw Toxic comment classification dataset.
  https://drive.google.com/file/d/18l6IwSqavnqtLQpVnrRqOugZf9XkhEAN/view?usp=sharing
- It consists of 5 categories:
  1. toxic
  2. severe_toxic
  3. obscene
  4. threat
  5. insult
  6. identity_hate
- The description of the files is given below:
  1. train.csv - the training set, contains comments with their binary labels
  2. test.csv - the test set, you must predict the toxicity probabilities for these comments. To deter hand labeling, the test set contains some comments which are not included in scoring.
  3. sample_submission.csv - a sample submission file in the correct format
  4. test_labels.csv - labels for the test data; value of -1 indicates it was not used for scoring;

2.2 Exercise

- Use TF-IDF and Word2Vec to represent the sentences.
- You will have to train your Word2Vec model seperately for this assignment,
- Use the Training data to train the SVM classifier using TF-IDF and Word2Vec representation.
- Use the Test data to evaluate your classifier
- Show results: Macro and Micro Precision, Macro and Micro Recall and Macro and Micro F1-Score. for both TF-IDF and Word2Vec

3 References

- https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_files.html
- https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
- https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-
- https://towardsdatascience.com/applying-machine-learning-to-classify-an-unsupervised-text-document
- https://www.kaggle.com/sandeepsingh3480/svm-toxic-comments-classification-challenge

4 Submission

- You have to submit your assignment in Jupyter notebook with proper comments and explanation of your approach.
- For each of your approach you have to show Micro precision, recall and F1 score for each class. And Macro precision, recall and F1 score.
- The submission deadline for this assignment in **23rd March 2021 at 11 PM**